French schwa and gradient cumulativity

Brian W. Smith
University of California, Berkeley

Joe Pater
University of Massachusetts Amherst

Abstract

We model the interaction of two phonological factors that condition French schwa alternations: schwa is more likely after two consonants; and schwa is more likely in the penultimate syllable. Using new data from a judgment study, we show that both factors play a role in schwa epenthesis and deletion, confirming previous impressionistic descriptions, and that the two factors interact cumulatively: they have a stronger effect together than alone. Treating each factor as a constraint, we find that their cumulative interaction in probabilistic space is better modeled with weighted rather than ranked constraints. To accomplish this, we characterize patterns of cumulativity in terms of how cumulativity affects probability. MaxEnt and Noisy HG can model the full range of cumulativity — sublinear, linear, and superlinear — while Stochastic OT can only model sublinear cumulativity. French schwa displays superlinear cumulativity, and as a result, the pattern is unobtainable in Stochastic OT. We find that the pattern of superlinearity is too extreme even for Noisy HG, leaving MaxEnt as the model with the closest fit to the experimental data.

1. Introduction

In his landmark study originally published in 1973, Dell (1985) provides a remarkably thorough description of the complex set of phonological factors conditioning the schwa-zero alternation in the relatively "standard" variety of Parisian French of which he is a native speaker, and proposes an analysis in terms of the phonological framework presented in Chomsky and Halle (1968). One of the central claims of his analysis is that both deletion of underlying schwa and epenthesis are involved in producing the surface distribution. Examples of deletion of underlying schwa are shown in (1a) and (1b), and a case of epenthesis is provided in (1c). French "schwa" is transcribed here as [œ].

1. Schwa deletion and epenthesis

(a)	/dœvrɛ/ → [dvrɛ]	Tu <u>devrais</u> partir.
(b)	/mœ/ → [m]	Tu <u>me</u> dois de l'argent.
(c)	/film/ → [filmœ]	un <u>film</u> danois

In all of these examples, the process applies variably: (1a, b) could be produced with a surface schwa, and (1c) without one. Although speech rate and speech register affect the probability of deletion in these contexts, according to Dell both variants are possible in what might be described as a neutral rate and register.

In this paper, we focus on the interaction of two phonological factors that affect the probability of schwa deletion and epenthesis. The first factor is whether a singleton consonant or a consonant cluster precedes the schwa. Deletion is less likely, and epenthesis more likely, when schwa is preceded by a cluster. Dell's rule of schwa deletion applies only when a single consonant precedes, as in the examples in (1a, b), and his rule of epenthesis applies only when a cluster ends the word (or morpheme), as in (1c). Dell's analysis abstracts from the fact that schwa deletion can also apply when a cluster precedes, as in (2a, b). Deletion almost certainly applies in these examples with lower probability than in (1a, b), but it is also possible for most speakers.

2. Examples of deletion in the CC_ context

(a) Il devrais partir. [il d_vʁɛ paʁtiɪ]

(b) Il me doit de l'argent. [il m_ dwa dœ laʁʒɑ̃]

A second factor that plays a role in conditioning the probability of both deletion and epenthesis is the position of the schwa in the prosodic phrase. Deletion is less likely, and epenthesis more likely, when the schwa is in penultimate position (see section 2 for references). For example, *film* is more likely to be followed by a schwa in *un film russe* [ɛ̃ filmœ ʁus] than *un film danois* [ɛ̃ filmœ danwa].

Examples like these raise both empirical and theoretical challenges. On the empirical side, the data on relative frequency of outcomes is harder to collect than are data on categorical differences. Single speaker intuitions and observations like those of Dell (1985) are invaluable as a starting point, but as we will show, they do not provide the fine-grained data needed to evaluate and compare probabilistic models. On the theoretical side, many phonological frameworks have no way to express the greater probability of the schwa-less pronunciation in (1a, b) than in (2a, b), let alone explain why particular contexts favor schwa. For example, the standard SPE framework adopted by Dell (1985) allows rules to apply categorically or optionally, but not with some specified probability. One could of course describe the patterns in a Variable Rules model (Labov 1969) by writing the conditioning factors into separate deletion and the epenthesis rules, but this model would not make particularly strong predictions. For example, there seems to be no reason that the preceding cluster could not increase the probability of deletion and decrease the probability of epenthesis, the opposite of observed facts.

Constraint-based models address both of these theoretical challenges. Such models allow a single factor, or constraint, to play a role across multiple processes, and as we will discuss below, there are several probabilistic constraint-based models that can generate degrees of optionality as required by the schwa data (see Coetzee & Pater 2011 for an overview of such models, and a comparison with Variable Rules). To model the two phonological factors discussed above, our analysis posits a constraint favoring schwa in penultimate position, and a constraint against coda clusters.

3. Constraints on schwa deletion and epenthesis

(a) PENULT=Ə
Effect: more schwa in _σ than in _σσ

(b) *CLUSTER
Effect: more schwa in CC_ than in C_

In French, these two constraints appear to interact *cumulatively*. Schwa is more likely to be realized in contexts where it's favored by both constraints, relative to contexts where it's favored by just one. Examples like *la terre se vend* (CC_σ) have greater probability of pronounced schwa than *le vin se vend* (C_σ) and *la terre se vend bien* (CC_σσ). In the table below, the rows distinguish schwas in penultimate *vs.* antepenultimate position, and the columns distinguish underlying schwas with a preceding singleton consonant *vs.* a cluster.

4. Examples of contexts with underlying schwa

	C_	CC_
_σ	le vin se vend [lœ vɛ̃ s_ vɑ̃]	la terre se vend [la tɛʁ s_ vɑ̃]
_σσ	le vin se vend bien [lœ vɛ̃ s_ vɑ̃ bjɛ̃]	la terre se vend bien [la tɛʁ s_ vɑ̃ bjɛ̃]

Although different constraint-based models of variation can model cumulativity, granting higher probability to a pronounced schwa in *la terre se vend* (CC_σ) than in the other examples, models differ in

the patterns they predict for the cells of table (4).

In this paper, we use these differences in predictions to compare three constraint-based models of variation in detail: Stochastic OT (Boersma 1997), Noisy Harmonic Grammar (Boersma & Pater 2016), and Maximum Entropy Grammar (MaxEnt; Goldwater & Johnson 2003). Stochastic OT can model the cumulative interaction of two constraints, as long as the constraints have relative similar values (Jäger & Rosenbach 2006). Cumulative constraint interaction is one of the major predictions of Harmonic Grammar (HG; Smolensky & Legendre 2006) whose weighted constraints produce "gang effects", and probabilistic variants of HG, such as Noisy HG and MaxEnt, can produce gradient cumulativity. In section 4, we characterize these differences in cumulativity in terms of how cumulativity affects probability: sublinearly, linearly, or superlinearly, and show that Stochastic OT produces only sublinear cumulativity, with the other theories having more subtle restrictions on the patterns they predict.

To compare the three frameworks, we report and model experimental data on French schwa, using judgments from multiple native speakers on the acceptability of pronounced schwa across contexts. Of the three models, MaxEnt provides the best fit for the pattern of superlinear cumulatively found in French schwa. Our results add to a growing body of work showing that weighted constraints provide a better fit to probabilistic natural language data than ranked constraints, particularly when it comes to cumulativity (Guy 1997; Benor & Levy 2006; Jäger & Rosenbach 2006; Zuraw & Hayes 2017)[1]. The French data also illustrate a prediction of weighted constraints that Zuraw and Hayes (2017) call across-the-board effects, which occur when a constraint has an effect on probabilities across contexts. In the case of French schwa, the effects of the conditioning factors are mirrored in both epenthesis and deletion contexts, modulo floor and ceiling effects. To our knowledge, this is the first model of French schwa to simultaneously address both probabilistic epenthesis and probabilistic deletion together.

The paper is structured as follows. Before presenting our experiment, we provide a brief review in section 2 of the two phonological factors conditioning of French schwa, and formalize these factors as phonological constraints. After the presentation of the experiment in section 3, we present a full model of the data in section 4, using the probabilities from the experiment to compare different constraint-based models of phonological variation.

2. Schwa epenthesis and deletion

In this section, we provide background on the two phonological factors, repeated in (5), which play a role in both schwa epenthesis and deletion.

5. Conditions on schwa epenthesis and deletion

   (a) The cluster factor: more schwa in CC_ than in C_
   (b) The phrase position factor: more schwa in _σ than in _σσ

The distinction we make between underlying and epenthetic schwas, adopted from Dell (1985), is as follows. Underlying schwas are morpheme-internal, such as the one in *devrais* in (1a), or in clitics, such as the one in *me* in (1b). Epenthetic schwas are found at the right edge of non-clitics, such as the schwa that appears after *film* in (1c). The justification for treating boundary schwas as epenthetic is the

[1] Two of these papers — Guy (1997) and Benor & Levy (2006) — compare ranked constraint models to logistic regression, which is equivalent to MaxEnt when there are two candidates per candidate set. All of the papers include Stochastic OT as a ranked constraint model, except Guy (1997), who only considers Anttila's (1997) model of partially ordered constraints.

alternation's productivity. Schwa can appear at *any* morpheme boundary, given the right phonological context. As shown in the examples below, schwa occurs at word boundaries (6a) and suffix boundaries (6b), even if there's no orthographic "e" (6c). In examples, we follow the notation of Dell (1985): in both orthography and phonetic transcription, obligatory schwas are underlined, optional schwas are in parentheses, and orthographic "e"s that are never pronounced are written ɸ.

6. Data from Dell (1985)

(a)	une veste rouge	[yn vɛstœ ʁuʒ]	(pg. 224)	*a red vest*
	cf. une vestɸ rouge et blanc	[yn vɛst ʁuʒ ɛ blã]	(pg. 224)	*a red and white vest*
(b)	exacte-ment	[ɛgzaktœ-mã]	(pg. 228)	*exactly*
	cf. massivɸ-ment	[masiv-mã]	(pg. 228)	*massively*
(c)	un short vert	[ɛ̃ ʃɔʁtœ vɛʁ]	(pg. 237)	*a green pair of shorts*

The underlying-epenthetic division we assume isn't universal. Dell (1985), for example, treats some of the alternating schwas above as underlying (consistent with orthography), and Côté & Morrison (2007) argue that schwas at clitic boundaries are epenthetic. As we show below, schwa in clitics and schwa at word boundaries surface at different rates, and the epenthesis-underlying distinction we assume provides a straightforward account of these differences.

2.1 The cluster factor

The cluster factor plays a role in both deletion and epenthesis. In both cases, schwa is more likely after two or more consonants than after one consonant. The examples below show this for deletion, while controlling for phrase position.

7. Deletion and the cluster factor (Dell 1985)

(a)	mange le gateau	[mãʒ lœ gato]	CCe σσ	(p. 229)
(b)	mangez l(e) gâteau	[mãʒɛ l(œ) gato]	C(e) σσ	(p. 229)
(c)	Jacques devrait partir	[ʒak dœvʁɛ paʁtiʁ]	CCe σσσ	(p. 228)
(d)	Henri d(e)vrait partir	[ãʁi d(œ)vʁɛ paʁtiʁ]	C(e) σσσ	(p. 228)

The number of preceding consonants also plays a role in epenthesis, as shown in (8). These data are judgments from Côté (2007).

8. Epenthesis and the cluster factor (Côté 2007)

(a)	la sect(e) partait	[la sɛkt(œ) paʁtɛ]	CC(e) σσ
(b)	l'Aztèquɸ partait	[l aztɛk paʁtɛ]	Cɸ σσ

Taken together, (7) and (8) show that epenthesis and deletion are affected by the cluster factor, but differ in their baseline rates of schwa: schwa is more likely to be pronounced when it's underlying. This can be seen by comparing (7a) *vs*. (8a) and (7b) *vs.* (8b).

In the constraint-based models that follow, we model the cluster factor with the constraint *COMPLEX, which militates against coda clusters.

9. *COMPLEX: Assign one violation for every coda cluster.

This constraint has been used in Optimality Theoretic accounts of French schwa deletion (Noske 1996; Tranel 2000) and reflects a syllable-based approach to French schwa, pursued in Morin (1974), Anderson (1982), and many others (for an overview, see Côté 2000: 87–90). The basic idea is that failing to produce a schwa in the context CC_C results in a marked coda cluster. For example, schwa is preferred in *la terre se vend bien* [la.tɛʁ.sœ.vɑ̃.bjɛ̃] because schwaless [la.tɛʁs.vɑ̃.bjɛ̃] contains the coda cluster [ʁs]. It's important to note that the constraint *COMPLEX is violable, and just one of many factors that

influence the likelihood of realizing of schwa.

The use of *COMPLEX here abstracts away from the sonority of coda clusters, which have well-documented effects on schwa realization. For example, Delattre (1951) describes schwa as more likely when it's preceded by two consonants of descending sonority (e.g. [ʁm_] in *ferm<u>e</u>ture*) than when it's preceded by two consonants of ascending sonority ([pʁ_] in *appr(e)nez*). These sonority effects aren't relevant in our model, since our experiment controls for cluster sonority.

2.2 Position

An effect of position has been observed since Léon (1966), who describes schwa as more likely to be pronounced in the penultimate syllable (see also Morin 1974, Dell 1985, Côté 2007). In epenthesis, the effect of the phrase position factor can only be observed after a cluster, since epenthesis reportedly never occurs after a single consonant.

10. Epenthesis and the phrase position factor (Morin 1974: 77)

(a)	le garde ment	[lœ gardœ mɑ̃]	CC<u>e</u> σ
(b)	le garde mentait	[lœ gaʁd(œ) mɑ̃tɛ]	CC(e) σσ

In deletion, an effect of the position factor can be observed after both clusters and singletons, word-internally and in clitics. The pair in (11) demonstrates that position plays a role when schwa is after two consonants.

11. Deletion shows an effect of position (Dell 1985: 231)

(a)	la terre s<u>e</u> vend	[la tɛʁ s<u>œ</u> vɑ̃]	CC<u>e</u> σ
(b)	la terre s(e) vend bien	[la tɛʁ s(œ) vɑ̃ bjɛ̃]	CC(e) σσ

The examples below show a similar effect of position after single consonants, although the effect is subtle. The schwas in (12) and (13) are described as optional, with schwa being more likely to be retained in the penultimate syllable.

12. *venez* in Dell (1985: 227)

v(e)nez ici	v(e)nez
[v(œ)ne isi]	[v(œ)ne]
← Less schwa	More schwa →

13. *ce* in Morin (1974: 77)

c(e) garçon	c(e) gars
[s(œ) gaʁsɔ̃]	[s(œ) ga]
← Less schwa	More schwa →

The phrase position effect described above distinguishes only between penult and non-penult positions. Other more nuanced effects of phrase position have been described, but these other effects are weaker and disputed. Côté (2007) reports that the prosodic context before schwa plays an additional role, as shown by pairs like *jette d<u>e</u> lortie* and *achète d(e) l'ortie,* in which schwa is preceded by one and two syllables, respectively. As noted by Côté, however, the effect of leftward context is arguably weaker than the effect of rightward context. Adding to the data in (12) and (13), both Dell and Morin claim that schwa becomes even less likely when it's followed by more syllables, as in *v(e)nez boire un verre* and *c(e) garçon-là*. Côté (2007) argues that the distinction between two following syllables and more than two following syllables is much weaker than the distinction between penult and non-penult. In support, Côté cites Lucci (1976), a corpus study on French schwa, which finds a distinction between one *vs.* two following syllables, but no distinction between two *vs.* three following syllables. In our experimental

items and model, the phrase position factor distinguishes only between penultimate schwas and non-penultimate schwas.

We model the phrase position factor with the constraint, PENULT = ə, which favors schwa when its followed by one syllable but not when it's followed by two (or more) syllables. This straightforwardly mirrors Léon's (1966) observation that schwa tends to be preserved or inserted in the penultimate syllable.[2]

14. PENULT = ə: Assign one violation if the penultimate syllable of the phonological phrase is a non-schwa vowel.

This constraint can be motivated on the basis of the prosodic structure of French and cross-linguistic tendencies in foot shape. Since French stress is phrase-final, it's been argued that every stress-assignment domain contains a single right-aligned iambic foot (Charette 1991: 146).[3] Support for an iambic analysis can be found in truncation, which typically creates final-stressed disyllables, e.g. *cinéma* → *ciné* [siˈne] (Scullen 1997).

Under an iambic analysis, the penultimate syllable of a phrase is special because it's the weak member of a foot, underlined below. The constraint PENULT = ə favors the realization of schwa in (15a), but is indifferent to schwa in (15b).

15. Footing of French phrases under the assumption of a single right-alighted iamb

(a)	*la terre se vend*	[la.tɛʁ.(sœ̲.ˈvɑ̃)$_{ft}$]
		[la.(tɛʁs̲.ˈvɑ̃)$_{ft}$]
(b)	*la terre se vend bien*	[la.tɛʁ.sœ.(vɑ̲̃.ˈbjɛ̃)$_{ft}$]
		[la.tɛʁs.(vɑ̲̃.ˈbjɛ̃)$_{ft}$]

Why should the weak member of a foot be schwa? This can be conceived of either as the result of foot-based sonority requirements or a drive for uneven iambs. In many languages, heads of feet contain vowels of higher sonority, while weak members of feet contain vowels of lower sonority (Kenstowicz 1994). Under the view that [œ] is the last sonorous vowel of French, it follows that the penultimate syllable will favor [œ] over other vowels. The claim that [œ] is the least sonorous vowel in French has been independently proposed in Tranel (2000: 68) to explain why it's the sole vowel to alternate with zero.

---

[2] Léon (1966: 122): "E tend à se maintenir à la pénultimième (type: garde-côte) ou à y apparaître (type: ours(e) blanc)."

[3] Charette (1991) assumes a single right-aligned foot to capture the position effect in schwa epenthesis. Under Charette's analysis in Government Phonology, the schwa in *gard<u>e</u>-boîte* is required to license the final consonant of *garde*. Without schwa, the output would be [gar.d.bwat], with an empty nucleus in the penultimate syllable. The position effect follows from the assumption that empty nuclei are generally allowed, but aren't licensed as the weak members of feet, resulting in epenthesis.

Another possibility is that the desire for schwa in the penultimate syllable comes from the interaction of stress and syllable weight. Across languages, iambic feet favor the shape of a light syllable followed by a heavy syllable (L˙H), and disfavor a heavy unstressed syllable (Hayes 1995). Under the assumption that codas contribute weight, realizing a schwa in the penultimate syllable (16a,b) ensures the penult is light, while producing a schwa outside of the penult (16c) offers no change in foot shape.

16. A weight-based account of PENULT = ə

(a)	*la terre se vend*	[la.tɛʁ.(sœ.ˈvɑ̃)$_{ft}$]	σσ[LL]
		[la.(tɛʁs.ˈvɑ̃)$_{ft}$]	σ[HL]
(b)	*la vin se vend*	[la.vɛ̃.(sœ.ˈvɑ̃)$_{ft}$]	σσ[LL]
		[la.(vɛ̃s.ˈvɑ̃)$_{ft}$]	σ[HL]
(c)	*la terre se vend bien*	[la.tɛʁ.sœ.(vɑ̃.ˈbjɛ̃)$_{ft}$]	σσσ[LL]
		[la.tɛʁs.(vɑ̃.ˈbjɛ̃)$_{ft}$]	σσ[LL]

For the experimental data we model, either justification of PENULT = ə, sonority or uneven iambs, is identical. Further data might be able to distinguish between them, but the question lies beyond the scope of this paper.[4]

2.3 Restrictions on schwa

One last restriction on schwa, which is relevant to our experimental design, is that schwa generally doesn't occur next to another vowel, even in contexts where the phrase position and cluster factors favor its pronunciation.

17. No schwa next to a vowel

(a)	l¢ homme	[lɔm] *[lœ ɔm]	*the man*
	cf. le gars	[lœ ga]	*the boy*
(b)	une ouvr¢-oeuf	[uvʁ œf] *[uvʁœ œf]	*egg opener*
	cf. ouvre-boîte	[uvʁœ bwat]	*can opener*

The main exception to this generalization is h-aspiré words, which phonetically begin with a vowel, but pattern in many ways as if they begin with a consonant. We set those aside here.

[4] A third possibility is that the phrase position factor is a result of stress clash avoidance, as argued in Mazolla (1992). Under this analysis, schwa in *la tèrre se vénd* avoids a clash between *tèrre* and *vénd*, while no clash is at stake in *la tèrre s(e) vend bién.* For our experimental items, a constraint like *CLASH will assign the same violations as PENULT = ə, assuming every lexical word has final stress. However, stress clash can't account for the phrase position factor's role in word-internal schwa epenthesis and deletion, as found in *exactement* (6b) or *venez* (12), since each word has only one stress. Interestingly, a clash-based analysis is considered and dismissed in previous analyses of schwa epenthesis, such as Charette (1991:167) and Côté (2007:4), in part because stress clash avoidance predicts a difference between C_σ and C_σσ, a difference which we do find in deletion contexts in our experiment.

2.4 Summary

The table below synthesizes judgments previously reported in the French schwa literature.

18. Both factors in deletion (word-internally and in clitics)

	C_	CC_
_σ	œ is optional	œ is obligatory
_σσ	œ is optional	œ is optional

19. Both factors in epenthesis (at non-clitic morpheme boundaries)

	C_	CC_
_σ	œ is forbidden	œ is obligatory
_σσ	œ is forbidden	œ is optional

A central claim of this paper is that both epenthesis and deletion are affected by the same phonological factors in the same ways, and should be modeled together. If deletion and epenthesis are conditioned by the same constraints, any phonological factor that favors schwa in epenthesis should also favor schwa in deletion (modulo ceiling or floor effects). We expect, for example, a difference between schwa deletion in C_σ *vs.* C_σσ, where both are described as optional, since a difference is observed between these contexts in epenthesis. We also expect the differences between deletion and epenthesis to be consistent across phonological contexts. If underlying schwa is more likely than epenthetic schwa after a single consonant, then underlying schwa should be more likely than epenthetic schwa after two consonants as well.

Three levels of optionality, as in (18) and (19), are insufficient to test these claims, or to fit and evaluate models of variation. For this reason, we present the results of an experiment to estimate the rate of schwa for each context.

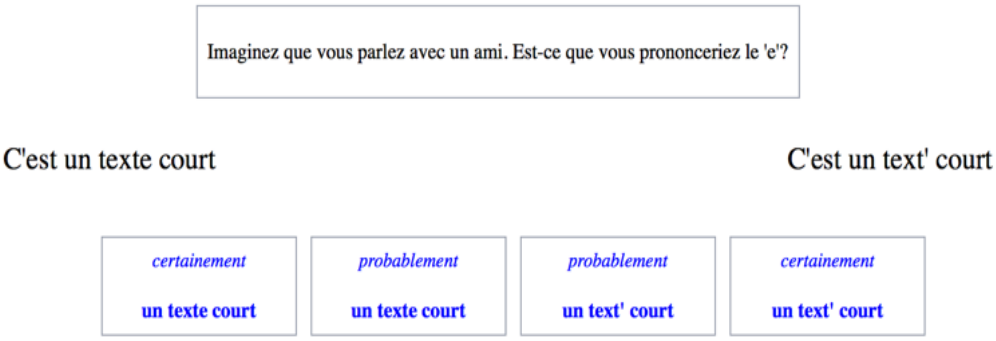
3. Experiment

3.1 Experimental design

We conducted the experiment over the internet, using IbexFarm (Drummond 2013). Participants were recruited by word of mouth through social media.

Task. The experimental task was two alternative forced choice. Participants were asked to imagine that they were speaking with a friend, and choose between two variants of a phrase: one with a pronounced schwa and one without a pronounced schwa. Choices were presented in French orthography. Pronounced schwa was indicated with an orthographic "e", and unpronounced schwa was indicated with an apostrophe, which is sometimes used to mark deleted schwas in songs to aid rhythmic parsing, or in some colloquially written words (e.g. *p'tit* for *petit*). For forms that didn't contain an "e" in the orthography, a pronounced schwa was indicated with an "e" in parentheses (e.g. *un lac(e) thai* vs. *un lac thai*). During a pre-experiment practice phase, participants received extra instructions for these forms. In addition to choosing between schwa and no schwa, participants indicated their confidence in the answer as *certainement* (definitely) or *probablement* (probably). A screen capture of the experiment in progress is in (20).

20. Screen capture of the experiment in progress



Design. The experiment followed a 2 x 2 x 2 factorial design, with 8 conditions. The three factors are below.

21. Factorial design

(a)	Cluster before schwa site	C_ vs. CC_
(b)	Position of schwa site	_σ vs. _σσ
(c)	Underlying or epenthetic schwa	schwa contained in clitic vs. non-clitic

The construction of items differed for underlying and epenthetic schwas. Underlying schwas were constructed according to the template below, consisting of a noun followed by a post-nominal adjective, with the site of the epenthetic schwa at the boundary between them.

22. **Noun** + **Adjective**
Noun: C-final or CC-final, all final Cs obstruents
Adjective: σ or σσ, all obstruent-initial

Nouns ended in either one or two consonants, and adjectives were one or two syllables long. All nouns in the experiment ended in obstruents, and all adjectives began with obstruents, controlling for the influence of sonority on the rate of schwa realization. Examples of the four epenthesis conditions are below. Each participant saw every noun and adjective only once.

23. Examples of epenthesis items

	C_	CC_
_σ	une bott(e) jaune [yn bɔt _ ʒon]	une vest(e) jaune [yn vɛst _ ʒon]
_σσ	une bott(e) chinoise [yn bɔt _ ʃinwaz]	une vest(e) chinoise [yn vɛst _ ʃinwaz]

Deletion items all consisted of the clitic *te*, the 2nd person object clitic, which we assume to be underlyingly /tə/. In these items, *te* was proceeded by a name and followed by a verb, e.g. *Maurice te cite* ('Maurice cites you'). The reason we used only one clitic is that different lexical items often differ in their rates of schwa deletion, a factor we wanted to control for.

24. **Name** + te + **Verb**
Name: C-final or V-final, all final Cs obstruents
Verb: σ (present) or σσ (imperfect), all obstruent-initial

The schwa in *te* is preceded by one consonant when the name is V-final, and two consonants when the name is C-final. Position of schwa was manipulated by using different tenses of verbs. In the present

tense, these verbs are monosyllabic. In the imperfect tense, the suffix *-ait* /-ɛ/ creates a disyllabic verb. Examples of the four deletion conditions are below. Each participant saw every name and verb lexeme only once.

25. Examples of deletion items

	C_	CC_
_σ	Eva t(e) choque [evat _ ʃok]	Maurice t(e) cite [moʁist _ sit]
_σσ	Eva t(e) choquait [evat _ ʃokɛ]	Maurice t(e) citait [moʁist _ sitɛ]

All items were checked for naturalness with three native speakers.

Each participant saw 6 items per condition, 24 for deletion and 24 for epenthesis, in addition to 30 fillers. Fillers consisted of tenses (past, future) and phonological environments that differed from the test items. Most importantly, some fillers contained phrases with schwa adjacent to vowels, which we used as catch trials. We excluded from analysis any participant who judged that schwa should definitely be pronounced next to a vowel. The design is summarized below.

26. 78 judgments per participant
    24 deletion: 6 per cell in (23), no name or verb repeated
    24 epenthesis: 6 per cell in (25), no adjective or noun repeated)
    20 fillers for deletion (e.g. Anna s(e) est levée)
    10 fillers for epenthesis (e.g. un iguan(e) solitaire)

3.2 Participants and exclusions

Participants were recruited online through word of mouth. There were 36 participants who self-identified as native French speakers from France. We excluded any participants who answered "definitely schwa" in catch trials once or more, leaving data for 27 participants.

3.3 Results

The proportion of schwa responses for both deletion and epenthesis contexts are presented in the tables and bar plot below. In parentheses, we include the range of the 95% confidence interval, specifically the Wilson score interval.

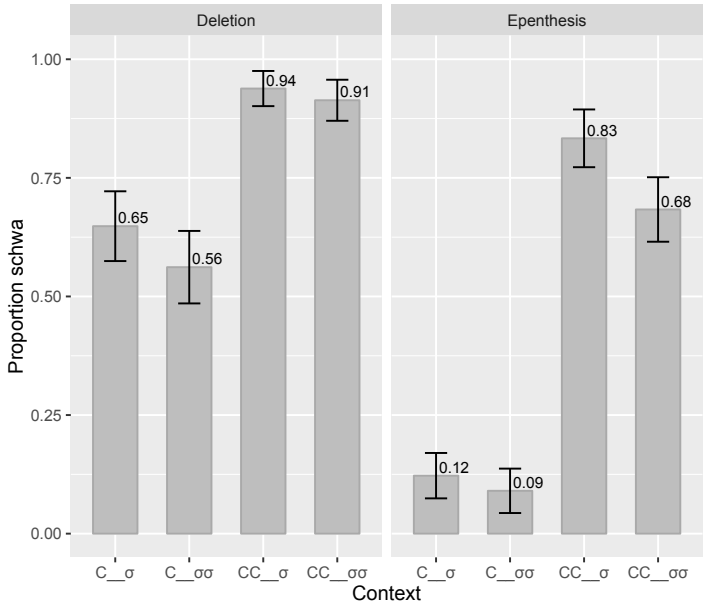
27. Deletion contexts: proportion schwa (Wilson score interval)

	C_	CC_
_σ	0.65 (0.57–0.72)	0.94 (0.89–0.97)
_σσ	0.56 (0.48–0.64)	0.91 (0.86–0.95)

28. Epenthesis contexts: proportion schwa (Wilson score interval)

	C_	CC_
_σ	0.12 (0.08–0.18)	0.83 (0.76–0.89)
_σσ	0.09 (0.05–0.14)	0.68 (0.61–0.75)

29. Barplot: whiskers show Wilson score intervals



Across all four phonological contexts, schwa is judged as better in deletion contexts than in epenthesis contexts. Schwa is also generally judged as better after two consonants than one consonant (the cluster factor), and better before one syllable than two syllables (the phrase position factor).

To evaluate the statistical significance and effect size of the factors, we fit a mixed effects logistic regression model in R (R Core Team 2017) using the package lme4 (Bates, Mächler, Bolker, & Walker 2015). The dependent variable in the model is the choice between schwa *vs.* no schwa. The model contains the fixed effects in the table in (30), each of which corresponds to an experimental condition, in addition to an interaction term for Stress × Seg. The model also contained a maximal random effects structure, with random intercepts for subject and item, and random slopes by subject for all of the fixed effects (including the interaction term).[6]

All of the categorical variables in the model were sum coded, as shown in the 'Coding' column in the table below. For each variable, the higher level (+1) is the context predicted to favor schwa.

30. Coding of fixed effects in model

Fixed effect	Level	Coding
Stress	_σ	+1
(position factor)	_σσ	–1
Seg	CC_	+1
(cluster factor)	C_	–1
Ep/Del	Deletion	+1
(epenthesis or deletion)	Epenthesis	–1

[6] The glmer equation in R: Schwa ~ EpDel + Stress * Seg + (1 | Item) + (1 + EpDel + Stress * Seg | Subject)

The fitted values for the model are shown in (31). The rightmost column, Pr>|Z|, shows p-values for Wald's test. A positive coefficient means the rate of schwa increases when the predictor is true (having a value of +1 in the table above), and decreases when the predictor is false. A negative coefficient means the rate of schwa decreases when the predictor is true.

31. Mixed effects model: logistic regression (positive = more schwa)

	Coefficient (β)	S.E.	Z	Pr>|Z|
(Intercept)	0.94	0.26		
Stress = _σ	0.31	0.11	2.70	<0.01
Seg = CC_	1.75	0.15	11.51	<0.001
Ep/Del = deletion	1.48	0.24	6.25	<0.001
Stress × Seg	– 0.06	0.11	0.55	0.59

All fixed effects were significant, except the interaction of Stress × Seg. As shown by the coefficient of Seg (β=1.75), the presence of a preceding cluster has the biggest effect on the realization of schwa: schwa is more likely after clusters than singletons. Schwa is also more likely in deletion contexts than epenthesis contexts (β=1.48), and more likely when followed by one syllable than when followed by two (β=0.31). Although the effect size of stress context is relatively small, it's significant in the model. The lack of significance for Stress × Seg shows that the effect of position is not limited to one segmental context (or vice versa). Both Stress and Seg exhibit independent effects on the likelihood of schwa realization.

4 Presentation of modeling results

In this section, we compare the ability of three models of variation to fit our experimental data: MaxEnt, Stochastic OT and Noisy HG. In the first section, we introduce the models by discussing some of the distributions that each one can generate for a subset of the French contexts, and some of the restrictions that each model places on the distributions it can generate relative to the other models. We then show how the models fare in fitting the actual French data. There has been some previous comparison of these theories (see Hayes and Macpherson 2016 and Pater 2016 and references therein); the following discussion draws in particular on Jäger and Rosenbach's (2006) comparison of Stochastic OT and MaxEnt, Pizzo's (2015) discussion of sublinearity in MaxEnt phonotactics, and Zuraw and Hayes' (2017) comparison of Noisy HG and MaxEnt with Stochastic OT.

4.1 The models

4.1.1 Constraint set and violation profiles

To illustrate how the models function, we will consider some distributions that they generate for a set of contexts in which schwa is supplied underlyingly, with the following constraints. For simplicity, we omit faithfulness constraints here, but include them below when needed.

32. *COMPLEX: Assign one violation for every coda cluster.

33. PENULT = ə: Assign one violation if the penultimate syllable of the phonological phrase is a non-schwa vowel.

34. NOSCHWA: Assign one violation for every schwa vowel in the output.

The contexts are those illustrated in the following table: the schwa is either in the penultimate syllable, or not, and it follows either a singleton, or a cluster.

35. Examples of contexts with underlying schwa

	C_	CC_
_σ	le vin se vend	la terre se vend
_σσ	le vin se vend bien	la terre se vend bien

We consider two candidates for each context: faithful realization of the schwa, and deletion. The tableau below shows violations for the two candidates in the context where both conditioning factors are relevant. Violations are marked with negative integers.

36. Constraint violations marked with negative integers

la terre se vend	NoSchwa	*Complex	Penult = ə
Deleted schwa: [la.tɛʁs.vɑ̃]		–1	–1
Pronounced schwa: [la.tɛʁ.sœ.vɑ̃]	–1		

The table in (37) uses the more compact representation of difference vectors, which result from subtracting the deletion candidate's violations from the faithful candidate's. Positive values indicate constraints that prefer schwa pronunciation, and negative values those that prefer deletion.

37. Difference vectors for constraint scores: negative values favor schwa deletion, positive favor schwa pronunciation

	NoSchwa	*Complex	Penult = ə
la terre se vend	–1	+1	+1
la terre se vend bien	–1	+1	0
le vin se vend	–1	0	+1
le vin se vend bien	–1	0	0

This representation clearly shows the trade-offs in constraint violations in each context. Faithful realization of the schwa always violates NoSchwa, and deletion always satisfies it, so all contexts have a value of –1 for NoSchwa, indicating a penalty for schwa pronunciation. This penalty trades off against a reward for schwa pronunciation that depends on the environment. In all of the models we consider, the probability of schwa pronunciation will always be the highest in the environment in which both *Complex and Penult = ə are relevant – the topmost row, and will always be the lowest in the environment in which neither is relevant – the bottom row. This sets these constraint-based models apart from a Variable Rules model. As we mentioned in the introduction, such a model could in principle make a schwa deletion rule apply with higher probability in any of the environments (see Coetzee and Pater 2011 for further related discussion). As we will shortly examine in detail, the three constraint-based models differ in exactly how rewards can accumulate in terms of differences in probability as we move up the move up the rows.

In Optimality Theory (OT: Prince & Smolensky 2004), schwa pronunciation is optimal *iff* a schwa-preferring constraint is ranked above NoSchwa. For example, given the ranking *Complex≫ NoSchwa≫Penult = ə, schwa pronunciation will be optimal in just the top two rows, in which *Complex prefers it. In a categorical version of Harmonic Grammar (HG; see Smolensky & Legendre 2006 and Pater 2016 and references therein), the optimal candidate is the one whose weighted sum of

constraint scores, or *Harmony*, is the highest. In terms of our difference vectors, schwa pronunciation is optimal when the sum of the difference scores, each times its constraints' weight, is above zero (see further Pater 2016). For example, if NOSCHWA had a weight of 3, and each of the other constraints had a weight of 2, schwa pronunciation would be optimal in only the top row, where the result of the just-described equation is 1. With these constraints, this gang effect pattern cannot be modeled in OT.

4.1.2 Probabilistic models of grammar

We now turn to the probabilistic variants of OT and HG that are our focus. In Maximum Entropy Grammar (MaxEnt; Goldwater & Johnson 2003), a probabilistic variant of HG, the probability of a candidate is proportional to the exponential of the weighted sum of violations. In terms of the difference vectors, the probability of the pronounced schwa is $\exp(n) / (1 + \exp(n))$, where n is the weighted sum of difference scores.[7] This means that the harmony difference between two candidates, *candidate a* minus *candidate-b*, is the log-odds of *candidate a*. A Harmony difference of 0 produces 0.5 probability, 1 → 0.73, 2 → 0.88, 3 → 0.95, 4 → 0.98, 5 → 0.99, and 6 → 1.0, all rounded to 2 decimal points. Negative Harmony differences equal one minus the positive value (–1 → 0.27, –2 → 0.12, –3 → 0.05 and so on). So, given the weights (3, 2, 2) from the previous paragraph, the probability of pronounced schwa would be 0.73 in the top row, 0.27 in each of the middle rows, and 0.05 in the bottom row.

Stochastic OT (Boersma 1997; Boersma & Hayes 2001) is a probabilistic variant of OT. Each constraint is given a real numbered value, but when the grammar is used to evaluate a candidate set, the numerical values are converted to an ordinal OT ranking. Variation occurs because the ranking values are perturbed by noise before conversion to ranking: each constraint value has a real number added to it that is sampled from a Gaussian distribution centered on zero (resampled for each constraint). As Jäger and Rosenbach (2006) point out, this model predicts greater probability in a gang effect context like the top row of the above table. To see this, consider the case when the constraints are tied in value (e.g. 1, 1, 1). In such a case, the probability of one constraint being ranked above another is 0.5, which is the probability of pronounced schwa in each of the middle rows. In the top row, the pronounced schwa is optimal if *either* *COMPLEX or PENULT = ə ranks above NOSCHWA, which obtains in 4/6 rankings, thus yielding a probability of 0.66.

Jäger and Rosenbach (2006) identify two differences between the patterns of gradient cumulativity that can be generated by Stochastic OT and MaxEnt. One is that if the two violations in the cumulative case come from a single constraint, in what they call counting cumulativity, Stochastic OT will not show an increase in probability under cumulativity. The other (p. 939) is an observation they attribute to Paul Boersma, that there are patterns of "strong" cumulativity that cannot be represented by Stochastic OT, but can be represented by MaxEnt. One such example is the MaxEnt pattern we mentioned above, in which the schwa pronunciation has 0.73 probability in the top row, but only 0.27 in the middle rows, and 0.05 in the bottom. In the classification scheme that we develop in what follows, this is a *superlinear* pattern, in that the probability increase from a single reward for schwa pronunciation (the middle rows) to two rewards (the top row) is greater than the sum of the increases gained by each of the single rewards on their own (bottom to middle rows): 0.46 *vs*. 0.44. As we show below, Stochastic OT

[7] The usual MaxEnt calculation for the probability of one of two candidates with Harmony H1 and H2 respectively is exp(H1) / exp(H1) + exp(H2). Because we have subtracted out the constraint scores for one of the candidates, its probability in the equation can be represented as exp(0) = 1. See Zuraw and Hayes (2017) for another derivation.

can only generate *sublinear* cumulativity.[8]

Noisy HG (Boersma & Pater 2016) is like Stochastic OT, except the values of the constraints are used in a weighted constraint evaluation of the candidate set. Like MaxEnt, it can generate superlinear cumultivity, though as we will see, the patterns the two models predict are not identical.

4.1.3 Detailed predictions of the models: sublinearity through superlinearity

To further explore the differences amongst these models, we will consider the patterns they each generate given particular values for the constraints. We first consider values of 1 for all of the constraints. For Noisy HG and Stochastic OT, the noise – the Standard Deviation of the Gaussian – is set to 0.2. For the Noisy HG model, any resulting negative weights were converted to zero (this is called Linear OT in Boersma and Weenink's 2017 Praat, which we used to explore these models). All probabilities in the table are rounded to two decimal points.

38. Proportion pronounced schwa in output distributions with constraints set to 1.

	Stochastic OT	Noisy HG	MaxEnt
la terre se vend	0.67	1	0.73
la terre se vend bien	0.5	0.5	0.5
le vin se vend	0.5	0.5	0.5
le vin se vend bien	0	0	0.27

As we have seen, for the middle two rows, one constraint prefers deletion (NOSCHWA), and one constraint prefers faithful schwa (*COMPLEX or PENULT = ə). With equal constraint values, all of the models grant equal probability to the two outcomes. The top row shows the cumulative effect of the two constraints that prefer the faithful candidate. We discussed above why Stochastic OT assigns a probability of 0.67 in this case. The 0.73 probability in MaxEnt arises because the two constraints preferring schwa pronunciation have a summed weight of 2, and NOSCHWA has a weight of 1, giving a difference of 1. In Noisy HG, a noise value of 0.2 has a very low probability of subverting the pre-noise preference for the faithful candidate by making the sum of the weights of *COMPLEX and PENULT = ə lower than NOSCHWA (less than 0.005, hence rounded to zero). In the final row, no constraint prefers the faithful candidate, and it has 0 probability in Stochastic OT. In Noisy HG, if a value of zero were sampled for NOSCHWA, the two candidates would be tied, and the tie would be broken with a random choice, which could yield the faithful candidate. The probability of this happening is less than 0.005. In MaxEnt, we again have a Harmony difference of 1 between the two candidates, but in this case the faithful candidate is the dispreferred one, which gets 0.27 probability.

The following table shows the result of increasing the constraint values to 2. For Stochastic OT, this has no effect. In Noisy HG, this results in a higher difference between the Harmonies of the candidates in the top and bottom rows, but the differences were already large enough with weights of 1 so that the noise of 0.2 had no perceptible effect given rounding. In MaxEnt, we see that there is now a higher probability for the faithful candidate in the top row, and for the deletion candidate in the bottom; these are the probabilities that result when the Harmony difference is 2. As we further increase the weight values, the probability of the faithful candidate will approach 1 in the top row, and 0 in the bottom.

[8] The formulation that Jäger and Rosenbach (2006) cite from Boersma is one in which the single reward cases each have probability less than ε, and the cumulative case has a probability > 1-ε. The example in the text shows that this definition is not identical to superlinearity, since it is not a "strong" case in Jäger and Rosenbach's (2006) terms.

Therefore, MaxEnt is capable of representing the more peaked distribution that Noisy HG produces with the current weights, at least to the degree of resolution we are examining.

39. Proportion pronounced schwa in output distributions with constraints set to 2.

	Stochastic OT	Noisy HG	MaxEnt
la terre se vend	0.67	1	0.88
la terre se vend bien	0.5	0.5	0.5
le vin se vend	0.5	0.5	0.5
le vin se vend bien	0	0	0.12

With this constraint set, Stochastic OT cannot produce the (1, 0.5, 0.5, 0) distribution over contexts. To get 0.5 for both of the middle contexts, NOSCHWA must have the same ranking value as both *COMPLEX and PENULT = ə, and in that case, the probability of pronounced schwa will always be 4/6 = 0.67.

It is also impossible for MaxEnt to represent the Stochastic OT distribution. To get near zero probability on faithful schwa in the bottom row, NOSCHWA must have a non-negligible weight. For instance, a weight of 5 will give it probability 0.007. To get 0.5 probability on faithful schwa for the middle rows, *COMPLEX and PENULT = ə must each have the same weight as NOSCHWA. Their summed weight will then give faithful schwa near 1 probability in the top row, failing to match the Stochastic OT value of 0.67. Noisy HG is also unable to match the Stochastic OT distribution: if the weights are small enough to allow NOSCHWA to overcome the cumulative effects of *COMPLEX and PENULT = ə with 0.67 probability when noise is added, a non-negligible number of faithful schwas will be produced in the bottom row (through random selection in a tie when both candidates have Harmony zero). For example, with the ranking values set to 0.2, the top row gets close to the Stochastic OT value at 0.72, and the middle rows are at 0.50, but the bottom is at 0.16. MaxEnt cannot match this Noisy HG distribution, for reasons we will now discuss.

In the examples we have looked at so far, cumulativity is weaker in Stochastic OT than in the weighted constraint theories. To measure cumulativity in probability space, we can consider the effect of a constraint on its own, versus its joint effect with another constraint. That is, in our tables, we can consider the differences in probability of schwa between the bottom and middle rows, which show the effects of each of *COMPLEX and PENULT = ə their own, versus the differences between middle and top, which show the additional probability given to schwa when both of the schwa-preferring constraints are relevant. In all of the weighted constraint examples we have looked at in the tables, cumulativity is linear: the difference between the bottom and middle rows is the same as the difference between the middle and top (this is only true with rounding for Noisy HG since the zero floor subverts true linearity). That is, the effect on probability of schwa of *COMPLEX or PENULT = ə is the same on its own as when it is added to the other constraint. Stochastic OT, on the other hand, displays sub-linear cumulativity: either constraint on its own increases the probability of schwa by 0.50, but the additional probability gained by adding the second constraint is only 0.17.

Stochastic OT can only represent sub-linear cumulativity. To see why, consider the six rankings of three constraints, one of which like NoSchwa disprefers some outcome X across the board (NoX) and two that like *Complex and Penult = ə prefer X in two partially overlapping environments ($+X_1$ and $+X_2$). In the table below, an X indicates that the outcome occurs in the environment specified in the column heading given the ranking in that row. The environments are those in which neither $+X_1$ nor $+X_2$ is relevant (Env. A), in which only one is (Env. B and Env. C), and in which both are (Env. D).

40. Illustration of Stochastic OT cumulativity

		Env. A	Env. B	Env. C	Env. D
a.	NoX >> $+X_1$ >> $+X_2$				
b.	NoX >> $+X_2$ >> $+X_1$				
c.	$+X_1$ >> NoX >> $+X_2$		X		X
d.	$+X_1$ >> $+X_2$ >> NoX		X	X	X
e.	$+X_2$ >> NoX >> +X1			X	X
f.	$+X_2$ >> $+X_1$ >> NoX		X	X	X

The difference in probability of X between Environment B, in which only $+X_1$ is relevant, and Environment A is the summed probability of rankings c., d., and f. The difference in probability of X between Environment C, in which only $+X_2$ is relevant, and the cumulative Environment D is just the probability of ranking c. Assuming that none of the rankings have zero probability, the probability difference of X will thus always be smaller between Environments C and D than between A and B: the contribution of $+X_1$ to the probability of X is greater on its own, than in conjunction with $+X_2$. The same logic applies to $+X_2$ on its own and jointly with $+X_1$.

MaxEnt and Noisy HG, on the other hand, can display a range of degrees of cumulativity, from sublinear through superlinear. As we have seen already, the degree of cumulativity is not completely free: neither MaxEnt nor Noisy HG could match Stochastic OT in the weakness of cumulativity in the examples in the above tables, when the probability in the middle rows was at 0.5. In this situation, MaxEnt is necessarily strictly linear. This can be understood based on Zuraw and Hayes' (2017) observation that the contribution of a given weighted constraint violation difference to probability forms a sigmoid that is steepest at 0.5, and which becomes shallower as we approach 0 and 1. In other words, its contribution is higher as we approach 0.5, and smaller as we approach 0 or 1. The contribution on either side of 0.5 is equal: if adding a violation difference increases probability from a baseline of 0.4 to 0.5, it will also increase probability from 0.5 to 0.6. This is the situation we have looked at in the tables thus far, and this explains why MaxEnt cannot match the Stochastic OT (0.67, 0.5, 0.5, 0) distribution in both tables, nor the Noisy HG (0.72, 0.5, 0.5, 0.12) distribution discussed in the text.

To escape the clutches of linearity in MaxEnt, we can change the probability of faithful schwa in the environment in which only one constraint applies. For example, if we give *COMPLEX and PENULT = Ə a higher value than NOSCHWA, such as 2 *vs.* 1 in the following table, the result of adding one of the constraints is a probability of higher than 0.5 as in the middle rows, and the effect of adding the other (difference with top row) will be smaller than its effect on its own (difference with bottom row). This is sublinear cumulativity, displayed here by all theories.

41. Proportion pronounced schwa in output distributions with NOSCHWA set to 1, and *COMPLEX and PENULT = Ə set to 2.

	Stochastic OT	Noisy HG	MaxEnt
la terre se vend	1	1	0.95
la terre se vend bien	1	1	0.73
le vin se vend	1	1	0.73
le vin se vend bien	0	0	0.27

MaxEnt can of course match the Stochastic OT and Noisy HG distributions to the degree of resolution we are examining. With the current constraint set, the MaxEnt distribution is completely out of reach of the other frameworks because the faithful schwa gets non-negligible probability in the bottom row, and it is harmonically bounded by deletion. To give them a chance to match it, we can add McCarthy and Prince's (1995) MAX to the constraint set, which assigns a violation to deletion in every context. To find weights, we used the learning procedure from the next section. In a typical run, Noisy HG was able to come close to the MaxEnt distribution with this larger constraint set (0.94, 0.73, 0.73, 0.25), but the Stochastic OT distribution remained fairly distant (0.89, 0.78, 0.78, 0.25), presumably because of its weaker cumulativity.

Finally, Noisy HG and MaxEnt can display superlinear cumulativity in probability differences, as shown in this last table, in which NOSCHWA is given a higher value than *COMPLEX and PENULT = Ə (again 2 *vs.* 1). In MaxEnt, we get predictable superlinearity when the result of adding a single constraint is probability less than 0.50. Here, the probability increase from the bottom to the middle rows is 0.15, and the increase from middle to top is 0.23.

42. Proportion pronounced schwa in output distributions with NOSCHWA set to 2, and *COMPLEX and PENULT = Ə set to 1.

	Stochastic OT	Noisy HG	MaxEnt
la terre se vend	0	0.5	0.5
la terre se vend bien	0	0	0.27
le vin se vend	0	0	0.27
le vin se vend bien	0	0	0.12

Since Stochastic OT is predictably sub-linear, superlinear patterns are predictably beyond its scope. MaxEnt and Noisy HG can of course model the Stochastic OT pattern by assigning NOSCHWA sufficient weight relative to the other constraints. With MaxEnt, we can model the NoisyHG pattern by scaling the weights used in the table (multiplying them by a constant), which will keep the top row at 0.50, and can bring the other rows as close to 0 as desired, and Noisy HG can in turn model the MaxEnt pattern, at least with the addition of MAX.

In sum, we have shown that each of the models can represent patterns of cumulativity that the others cannot. This means that we should be able to test them in their relative ability to match natural language cumulativity. The biggest difference amongst the models appears to be Stochastic OT's weaker cumulativity with respect to the other two: it is always sub-linear. MaxEnt's degree of cumulativity, sub-linear, linear, or superlinear, was shown to be related to where the effect of a single competing

constraint lands in probability space, below 0.50, at 0.50, or above. Noisy HG's degree of cumulativity is less predictable in that it can model sub-linear patterns out of reach of MaxEnt, and in that respect, seems like it falls between the two other theories, as might be expected as it combines Stochastic OT's noise with MaxEnt's weighted evaluation.

4.2 Models fit to French data

Along with cases of underlying schwa discussed in the previous section, our judgment experiment examined four parallel epenthesis contexts, illustrated in the following table, with the potential schwas underlined.

43. Examples of epenthetic schwa contexts

	C_	CC_
_σ	la bott<u>e</u> jaune	mets ta vest<u>e</u> rouge
_σσ	la bott<u>e</u> chinoise	mets ta vest<u>e</u> marron

We assume that the vowels in these cases are not underlying, but are supplied through epenthesis. In the contexts in the rightmost column, the epenthetic schwa avoids a coda cluster, and in those in the top row, it provides a penultimate schwa.

The experimental grand means of pronounced schwa choice are repeated in the two tables below, rounded to three decimal points (more precise values were used for finding constraint values). In both tables, the lowest rate of schwa is in the bottom-left cell, where the schwa is in the antepenultimate syllable with only a single preceding consonant, and the highest rate is in the upper-right cell, where schwa is in penultimate syllable with two preceding consonants. Intermediate values obtain when only the constraint against clusters is relevant (bottom-right cell), or the constraint against singletons (top-left cell). The presence of an underlying vowel leads to a higher rate of schwa in all contexts.

44. Experimental results (*p.* of pronounced vowel)

	Underlying			Epenthetic	
	C_	CC_		C_	CC_
_σ	0.648	0.938	_σ	0.122	0.833
_σσ	0.562	0.914	_σσ	0.090	0.683

The constraint set for these models includes the three markedness constraints introduced in the last section for the deletion cases: NoSchwa disprefers schwa across the board, and *Complex and Penult = ə prefer it in the environments shown in the rightmost columns and the top rows of our tables respectively. The faithfulness constraint Max prefers the pronounced schwa when it is underlying, and Dep prefers its absence when it would need to be supplied through epenthesis (see McCarthy & Prince 1995 on Max and Dep). We also include NoCoda because schwa is preferred by none of the other constraints in the epenthesis context represented by bottom-left cell, so Stochastic OT would be unable to grant it any probability, and would be unable to match the empirical value of 0.090. The preferences of the full constraint set for both underlying and epenthetic schwa are shown in the following table. With this constraint set any of the three frameworks can match the data in an individual cell of the table in (45) to arbitrary precision, and they can also get the general pattern of cumulative constraint interactions. The question is how closely they can fit the overall pattern.

45. Difference vectors for constraint scores: negative values favor schwa deletion, positive favor schwa pronunciation

	NoSchwa	*Complex	Penult = ə	Max	Dep	NoCoda
la terre se vend	–1	+1	+1	+1	0	0
la terre se vend bien	–1	+1	0	+1	0	0
le vin se vend	–1	0	+1	+1	0	+1
le vin se vend bien	–1	0	0	+1	0	+1
mets ta veste rouge	–1	+1	+1	0	–1	0
mets ta veste marron	–1	+1	0	0	–1	0
la botte jaune	–1	0	+1	0	–1	+1
la botte chinoise	–1	0	0	0	–1	+1

We first present a MaxEnt model whose weights were obtained by using a batch learner (Staubs 2011) that incorporates an optimization algorithm that finds weights that minimize the difference between the training data and the model predictions, in terms of Kullback–Leibler divergence (Kullback & Leibler 1951). This is an implementation of the same general approach to MaxEnt grammar and learning that is presented in Goldwater and Johnson (2003) and Wilson (2006) (as well as Hayes & Wilson 2008, though their model defines a probability distribution over all possible words, rather than over a set of candidates for a given UR). The optimization algorithm was L-BFGS-B (Byrd, Lu, Nocedal, & Zhu 1995) as implemented in R (Bates et al. 2015). The weights were constrained to be above zero, and a gaussian prior with variance 100,000 was imposed (the prior seemed to have no effect, as a weaker prior did not change the solution). The tables below show the predicted probabilities for pronounced schwa in each of the 8 environments, as well as the difference with respect to the empirical data in the rows labeled error (positive values indicate that the predicted value is too high, negative too low). The sum of absolute differences for this MaxEnt model with respect to the empirical data is 0.253 (mean over contexts = 0.032; we present SSE and K-L divergence in the summary table at the end of this section).

46. MaxEnt predictions after batch training (probability of pronounced vowel)

	Underlying			Epenthetic	
	C_	CC_		C_	CC_
_σ	0.633	0.967	_σ	0.167	0.775
error	–0.015	0.029	error	0.045	–0.058
_σσ	0.514	0.948	_σσ	0.109	0.678
error	–0.048	0.034	error	0.019	–0.005

The constraint weights producing these probabilities are shown in the following table. As mentioned in the previous section, the probabilities result from the formula $\exp(n) / (1 + \exp(n))$, where n is the weighted sum of difference scores. For the *la botte chinoise* type of epenthetic schwa, whose probability is 0.109, the weighted sum is the negative of the weights of NoSchwa and Dep, plus the weight of NoCoda: –1.015 + –1.084 + 0 = –2.099. The corresponding underlying schwa type, *le vin se vend*, differs in the absence of the negative contribution of Dep, and presence of the positive contribution of Max, thus leading to a higher baseline probability of schwa in the "Underlying" table in (46).

47. MaxEnt constraint weights after batch training

*COMPLEX	2.845
DEP	1.084
MAX	1.069
NOSCHWA	1.015
PENULT = ə	0.490
NOCODA	0.000

The contribution of the high weighted *COMPLEX is seen in the probability differences between the columns in each of the "Underlying" and "Epenthetic" tables in (46), while contribution of the somewhat lower weighted PENULT = ə is seen in the probability differences between rows.[9] As discussed in the previous section, the function relating weight differences to probability differences is a sigmoid centered at 0.50 probability. Therefore, the highest possible contribution of a weight difference is when the midpoint between the probability where the constraint doesn't apply, and the probability where it does apply, is 0.50. Thus, the greatest contribution of the *COMPLEX constraint is in the penultimate epenthetic context, where it yields a probability increase of 0.608 (0.775 – 0.167), and the midpoint is closest to 0.50 (0.471). This is in line with the empirical differences, where this context has the highest difference between preceding singleton and cluster. One might think that to get a greater difference between the top two cells in the Epenthetic table than in the Underlying table one would need a separate constraint, but in fact, this follows in the MaxEnt model from the difference in the baseline probability value in each case. Since the baseline probability in the Underlying case is the singleton probability of 0.633, the MaxEnt model is predicted to yield a smaller probability increase in the cluster. It is worth noting, though, that the MaxEnt model winds up producing a slightly smaller difference between the columns than in the empirical data for the Epenthetic table, and a slightly larger difference for the Underlying table.

PENULT = ə has its greatest effect on probability differences in the realization of underlying schwa in the C_ environment (0.633 – 0.514 = 0.119), again because the midpoint is the closest to 0.50. This fits the empirical data in terms of producing a greater effect for PENULT = ə in the singleton than in the cluster environment within the Underlying table, and also in terms of producing a greater effect for PENULT = ə in singletons in the Underlying case than in the Epenthetic. One subtle mismatch with the empirical data is that the greatest effect for PENULT = ə is in fact in the cluster environment of the Epenthetic table (the rightmost column). The MaxEnt model cannot match this because the baseline in that case is further away from 0.50.

To obtain fitted models for Stochastic OT and Noisy HG, we must use on-line learners; no batch approaches are available because it is computationally costly to calculate or estimate model predicted probabilities in those frameworks. In on-line learning, the learner receives a single piece of data at each learning step and uses the grammar to generate a prediction just for that datum, updating the constraint values if the learning datum and the prediction mismatch. Conveniently, it is possible to conduct on-line learning in a nearly identical way across the three frameworks. For MaxEnt, the on-line method is referred to as Stochastic Gradient Ascent (Jäger 2007), and in applying it to Noisy HG, Boersma and Pater (2016) call it the Harmonic Grammar Gradual Learning Algorithm (HG-GLA). The weights are updated by the difference in violation vectors between the learner's prediction and the learning datum, scaled by a learning rate, or plasticity. In Stochastic OT's GLA, constraints preferring the correct learn-

[9] Because NOCODa applies only in the singleton contexts, the effective value of *COMPLEX is diminished by the weight of NOCODA, but NOCODA here has a zero weight.

ing datum are promoted by the plasticity amount, and those preferring the learner's own incorrect prediction are demoted. When the differences between the candidate vectors are always zero or one, as in our examples (see table 41), the HG-GLA and the OT-GLA are identical.

The learning simulations were conducted in Praat (Boersma & Weenink 2017). Constraints were given an initial value of 2, and the plasticity was set to 0.1. The learners received 100,000 samples from the target distributions. These distributions were the experimental results in section 3, with equal probability given to each off the 8 contexts. The learner then received 3 more sets of 100,000 samples of data, with the plasticity set at 0.01, 0.001 and 0.0001 respectively. This training regime is based on the Praat defaults, but with an initial weight value of 2 rather than 10 so as to get comparable results across the frameworks, and with a correspondingly lower initial plasticity. The noise for Stochastic OT and Noisy HG was set at 0.2, rather than the Praat default of 2, because of the lower initial weight and plasticity. We conducted 20 runs for each model.

The MaxEnt model trained on-line predicts distributions very similar to those of the model trained in a batch fashion. The following table shows the results from the model that provides the closest fit to the data, with a sum of absolute differences of 0.240 (mean over contexts = 0.030). The 20 runs had an average summed absolute difference of 0.256 (mean 0.032), with a maximum of 0.269 (mean 0.034).

48. MaxEnt predictions after on-line training (*p.* of pronounced vowel)

	Underlying			Epenthetic	
	C_	CC_		C_	CC_
_σ	0.637	0.968	_σ	0.166	0.778
error	-0.011	0.030	error	0.043	-0.056
_σσ	0.518	0.950	_σσ	0.109	0.682
error	-0.043	0.036	error	0.019	-0.001

The weights producing that distribution are somewhat different from those for the batch model, but we again have a relatively high weight for *COMPLEX, and a relatively low weight for PENULT = ə.

49. MaxEnt constraint weights after on-line training

*COMPLEX	3.532
NOSCHWA	1.798
MAX	1.184
DEP	0.982
NOCODA	0.670
PENULT = ə	0.502

The predictions of the best fitting Stochastic OT model are shown in the following table. The sum of absolute differences with respect to the empirical data is higher than the best MaxEnt model, 0.299 (mean 0.037). The average sums of absolute differences over 20 runs was 0.330 (mean 0.041), and the maximum was 0.381 (mean 0.048). The distributions of these error measures for the MaxEnt models and the Stochastic OT models are non-overlapping: the worst fitting of the 20 on-line MaxEnt models had less error than the best fitting of the Stochastic OT models. We'll show shortly that this holds for other ways of measuring error as well.

50. Stochastic OT predictions (*p.* of pronounced vowel)

	Underlying			Epenthetic	
	C_	CC_		C_	CC_
_σ	0.648	0.914	_σ	0.169	0.769
error	0.000	-0.025	error	0.047	-0.064
_σσ	0.567	0.907	_σσ	0.109	0.756
error	0.005	-0.006	error	0.012	0.073

Like the MaxEnt models, the Stochastic OT predictions get the general pattern of cumulative constraint interactions, and the individual fits are sometimes even somewhat better. The bulk of the error is in the rightmost column of the Epenthetic table: the values of the two rows are too close together with respect to the empirical data, which means the effect of PENULT = ə in the cumulative interaction with *COMPLEX is too weak. In the empirical data, the cumulative effect of *COMPLEX is superlinear: there is a 0.032 difference in the C_ context, and a 0.150 difference in the CC_ context. As discussed in the last section, Stochastic OT produces cumulative interactions that are predictably sub-linear in probability space, here leading to a gross mismatch with the empirical data, which show a 0.060 difference in the C_ context, and 0.013 in CC_ .

The Stochastic OT constraint values producing this distribution are shown in the following table. In contrast with the MaxEnt values, the Stochastic OT values are much closer together. This is because variation, and the consequent cumulativity, requires constraints to be relatively close in value so that their ranking will vary across samples from the noise distribution. Nonetheless, we see the same general pattern of *COMPLEX having a higher value than PENULT = ə.

51. Stochastic OT constraint values

*COMPLEX	2.402
DEP	2.144
MAX	2.097
NOSCHWA	2.047
PENULT = ə	1.977
NOCODA	1.551

The final set of predictions are those of the best fitting Noisy HG model. The sum of absolute differences with respect to the empirical data is comparable to the best Stochastic OT model, 0.295 (mean 0.037). The average over 20 runs was also similar, 0.327 (mean 0.041), as was the maximum, 0.381 (mean 0.0486). The distribution of error over the eight contexts was somewhat different; the best fitting model is again typical.

52. Noisy HG predictions (*p.* of pronounced vowel)

	Underlying			Epenthetic	
	C_	CC_		C_	CC_
_σ	0.634	0.977	_σ	0.195	0.766
error	-0.014	0.038	error	0.072	-0.067
_σσ	0.527	0.963	_σσ	0.107	0.69
error	-0.035	0.050	error	0.016	0.002

The Noisy HG model succeeds in getting a greater spread than Stochastic OT between the contexts in the CC_ column of the Epenthetic table, in this respect mimicking MaxEnt, and approaching the empirical spread. In doing this, though, it also creates a greater spread between the values in the C_ column than motivated by the empirical data. Here Noisy HG is producing a slightly sublinear pattern: the effect of PENULT = ə on the probability is a 0.088 difference on its own (penultimate column), and 0.081 in conjunction with *COMPLEX (rightmost). In this respect, it is intermediate between the superlinear pattern of MaxEnt, and the highly sublinear pattern of Stochastic OT. Noisy HG patterns like MaxEnt in giving both contexts in the CC_ column of the Underlying table too much probability of pronounced schwa, and both contexts of C_ too little; these models are not quite fitting the extent to which *COMPLEX has a greater effect in the Epenthetic contexts.

The weights producing the Noisy HG distribution are given in (53). As in MaxEnt, the additive nature of constraint interaction in this weighted constraint model allows constraints with even small weights to have an effect on the outcome. Again, the greater effect of *COMPLEX than PENULT = ə seen in the probability distributions is reflected in the weights, even allowing for the effect of NOCODA in singleton contexts.

53. Noisy HG constraint weights

*COMPLEX	2.299
NOSCHWA	1.955
NOCODA	1.746
MAX	0.211
DEP	0.166
PENULT = ə	0.034

In sum, all three models – MaxEnt, Noisy HG and Stochastic OT – were able to capture the general pattern of cumulative constraint interaction seen in the empirical data, and provided reasonable fits to the attested values. The MaxEnt model did slightly better than the other models, and in comparison with Stochastic OT, at least some of that success is attributable to its ability to produce superlinear cumulativity in probability space.

Our comparisons of models' fit to the empirical data have thus far been made in terms of differences in raw probability. There are other ways of measuring fit, and one might wonder whether the outcome is different using other metrics. In the following table, we provide the mean, best and worst fits for each model in terms sum of squared error and Kullback-Lieber divergence, and also repeat the

absolute error values reported in the text.[10] In all cases, MaxEnt had consistently lower error than the other models. When error is measured in terms of SSE or K-L Divergence, the Noisy HG values are lower than those for Stochastic OT, and the MaxEnt vs. Stochastic OT difference is enhanced, because the error in the Stochastic OT predictions is concentrated in just two of the contexts (the _CC column of the Epenthetic table).

54. Error for each model, fitted on experimental data

	Absolute Error			Sum of Squared Error			K-L Divergence		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
Stochastic OT	0.330	0.299	0.381	0.043	0.037	0.052	0.086	0.064	0.112
NoisyHG	0.327	0.295	0.371	0.035	0.031	0.045	0.035	0.034	0.037
MaxEnt	0.256	0.240	0.269	0.021	0.019	0.023	0.020	0.020	0.021

5. Conclusion

In this paper, we described and modeled the interaction of two phonological factors that condition French schwa alternations: schwa is more likely after two consonants (the cluster factor) and in the penultimate syllable (the phrase position factor). Each of these factors has been identified in the literature on French schwa, but their interaction in probability space hasn't been previously described or formalized. Using data from a judgment study, we showed that the two factors interact superlinearly, and that both factors play a role in schwa epenthesis and deletion. Treating each factor as a constraint, we found that MaxEnt is the best model of the constraints' interaction. MaxEnt and Noisy HG can model the full range of cumulativity — sublinear, linear, and superlinear — while Stochastic OT can only model sublinear cumulativity. MaxEnt's advantage over the other two models comes from the fact that French schwa displays superlinear cumulativity, which is unobtainable in Stochastic OT and, in this case, too extreme for Noisy HG.

[10] Absolute error was calculated with respect to the probability of schwa in each context. Sum of squared error and K-L divergence were calculated over the probability of each of schwa and no-schwa. K-L divergence is formulated to be calculated over entire probability distributions. If SSE were calculated over just probability of schwa, the value would be half of that reported, and if absolute error were calculated for both schwa and no-schwa, it would double.

References

Anderson, Stephen R. 1982. The analysis of French schwa: or how to get something for nothing. *Language*. 58. 534–573.

Anttila, Arto. 1997. Deriving variation from grammar. In F. Hinskens, R. van Hout, & W. L. Wetzels (eds.), *Variation, Change, and Phonological Theory*, 35–68. Amsterdam: John Benjamins.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*. 67(1). 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Benor, Sarah & Roger Levy. 2006. The chicken or the egg? A probabilistic analysis of English binomials. *Language*. 82(2). 233–278.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*. 21. 43–58.

Boersma, Paul & Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*. 32(1). 45–86.

Boersma, Paul & Joe Pater. 2016. Convergence properties of a gradual learning algorithm for Harmonic Grammar. In J. McCarthy & J. Pater (eds.), *Harmonic Grammar and Harmonic Serialism*,. London: Equinox Press.

Boersma, Paul & David Weenink. 2017. Praat: doing phonetics by computer (Version Version 6.0.29).

Byrd, Richard H., Peihuang Lu, Jorge Nocedal & Ciyou Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*. 16(5). 1190–1208.

Charette, Monik. 1991. *Conditions on phonological government*. Cambridge University Press.

Chomsky, Noam & Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper and Row.

Coetzee, Andries W. & Joe Pater. 2011. The place of variation in phonological theory. In J. A. Goldsmith, J. Riggle, & A. C. Yu (eds.), *Handbook of Phonological theory*, 2nd ed., 401–434. Wiley-Blackwell.

Côté, Marie-Hélène. 2000. *Consonant Cluster Phonotactics: A Perceptual Approach* (Doctoral dissertation). MIT.

Côté, Marie-Hélène. 2007. Rhythmic constraints on the distribution of schwa in French. In V. Camacho J.Deprez, N. Flores, & L. Sanchez (eds.), *Proceedings of LSRL 36*,. Amsterdam: John Benjamins.

Côté, Marie-Hélène & Geofferey Stewart Morrison. 2007. The nature of the schwa-zero alternation in French clitics: experimental and non-experimental evidence. *Journal of French Language Studies*. 17. 159–186.

Delattre, Pierre. 1951. *Principes de phonétique française à l'usage des étudiants anglo-américans*. École Française d'Êté Middlebury College.

Dell, François. 1985. *Les règles et les sons*. Paris: Hermann.

Goldwater, Sharon & Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy

model. In J. Spenader, A. Eriksson, & Ö. Dahl (eds.), *Proceedings of the Workshop on Variation within Optimality theory*, 111–120. Stockholm University.

Guy, Gregory R. 1997. Violable is variable: Optimality Theory and linguistic variation. *Language Variation and Change*. 9. 333–347.

Hayes, Bruce. 1995. *Metrical Stress Theory: Principles and Case Studies*. Chicago: The University of Chicago Press.

Hayes, Bruce & Colin Wilson. 2008. A Maximum Entropy Model of Phonotactics and Phonotactic Learning. *Linguistic Inquiry*. 39(3). 379–440.

Jäger, Gerhard. 2007. Maximum entropy models and stochastic Optimality Theory. *Architectures, Rules, and Preferences: Variations on Themes by Joan W. Bresnan. Stanford: CSLI*. 467. 479.

Jäger, Gerhard & Anette Rosenbach. 2006. The winner takes it all - almost. Cumulativity in grammatical variation. *Linguistics*. 44(5). 937–971.

Kenstowicz, Michael. 1994. *Sonority-Driven Stress*.

Kullback, Solomon & Richard A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*. 22(1). 79–86.

Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language*. 45. 715–762.

Léon, Pierre R. 1966. Apparition, maintien et chute du" e" caduc. *La Linguistique*. 2(Fasc. 2). 111–122.

Lucci, Vincent. 1976. Le mécanisme du 'E' muet dans différentes formes de français parlé. *La Linguistique*. 12(2). 87–104.

McCarthy, John J. & Alan Prince. 1995. Faithfulness and Reduplicative Identity. In J. Beckman, L. Walsh Dickey, & S. Urbanczyk (eds.), *University of Massachusetts Occasional Papers in Linguistics 18*, 249–384. Amherst, Mass.: GLSA Publications.

McPherson, Laura & Bruce Hayes. 2016. Relating application frequency to morphological structure: the case of Tommo So vowel harmony. *Phonology*. 33(01). 125–167.

Morin, Yves-Charles. 1974. Règles phonologiques à domaine indéterminé: Chute du cheva en français. *Cahiers de Linguistique de l'Université Du Québec*. 4. 69–88.

Pater, Joe. 2016. Universal grammar with weighted constraints. *Harmonic Grammar and Harmonic Serialism*.

Pizzo, Presley. 2015. *Investigating Properties of Phonotactic Knowledge Through Web-Based Experimentation*. University of Massachusetts Amherst.

Prince, Alan & Paul Smolensky. 2004. *Optimality theory : constraint interaction in generative grammar*. Malden, MA: Blackwell Pub.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Scullen, Mary Ellen. 1997. *French prosodic morphology: a unified account*. Indiana University Linguistics Club Publications.

Smolensky, Paul & Géraldine Legendre. 2006. *The Harmonic Mind: From Neural Computation to Op-*

timality-Theoretic Grammar. Cambridge, MA: MIT Press.

Staubs, Robert. 2011. *HG in R (hgR). Software package*. Amherst, MA: University of Massachusetts Amherst. Retrieved from Software available at http://blogs.umass.edu/hgr/hg-in-r

Tranel, Bernard. 2000. Aspects de la phonologie du français et la théorie de l'optimalité. *Langue Française*. (126). 39–72.

Wilson, Colin. 2006. Learning Phonology with Substantive Bias: An Experimental and Computational Study of Velar Palatalization. *Cognitive Science*. 30(5). 945–982.

Zuraw, Kie & Bruce Hayes. 2017. Intersecting constraint families: an argument for Harmonic Grammar. *Language*. 93(3).

Zuraw, Kie & Bruce Hayes. 2017. Intersecting constraint families: an argument for Harmonic Grammar. *Language*. 93(3). 497–548.